IEA
International Epidemiological Association

International Journal of Epidemiology, 2023, 214–226
https://doi.org/10.1093/ije/dyac130
Advance Access Publication Date: 24 June 2022
Original article

OXFORD

Methods

Linkage of multiple electronic health record datasets using a 'spine linkage' approach compared with all 'pairwise linkages'

Helen A Blake ,[1,2]* Linda D Sharples ,[3] Katie Harron ,[4] Jan H van der Meulen[1,2] and Kate Walker[1,2]

[1]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK, [2]Clinical Effectiveness Unit, Royal College of Surgeons of England, London, UK, [3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK and [4]Population, Policy & Practice Department, University College London (UCL) Great Ormond Street Institute of Child Health, UCL, London, UK

*Corresponding author. Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15–17 Tavistock Place, London, WC1H 9SH, UK. E-mail: helen.blake@lshtm.ac.uk

Received 9 January 2022; Editorial decision 13 May 2022; Accepted 13 June 2022

Abstract

Background: Methods for linking records between two datasets are well established. However, guidance is needed for linking more than two datasets. Using all 'pairwise linkages'—linking each dataset to every other dataset—is the most inclusive, but resource-intensive, approach. The 'spine' approach links each dataset to a designated 'spine dataset', reducing the number of linkages, but potentially reducing linkage quality.

Methods: We compared the pairwise and spine linkage approaches using real-world data on patients undergoing emergency bowel cancer surgery between 31 October 2013 and 30 April 2018. We linked an administrative hospital dataset (Hospital Episode Statistics; HES) capturing patients admitted to hospitals in England, and two clinical datasets comprising patients diagnosed with bowel cancer and patients undergoing emergency bowel surgery.

Results: The spine linkage approach, with HES as the spine dataset, created an analysis cohort of 15 826 patients, equating to 98.3% of the 16 100 patients identified using the pairwise linkage approach. There were no systematic differences in patient characteristics between these analysis cohorts. Associations of patient and tumour characteristics with mortality, complications and length of stay were not sensitive to the linkage approach. When eligibility criteria were applied before linkage, spine linkage included 14 509 patients (90.0% compared with pairwise linkage).

Conclusion: Spine linkage can be used as an efficient alternative to pairwise linkage if case ascertainment in the spine dataset and data quality of linkage variables are high. These aspects should be systematically evaluated in the nominated spine dataset before spine linkage is used to create the analysis cohort.

Key words: Record linkage, pairwise linkage, spine linkage approach, electronic health records

Key Messages

- The spine approach to linking multiple datasets can reduce the number of linkages required and thus is more time-efficient, resource-efficient and cost-efficient compared with obtaining all pairwise linkages.
- All methodological decisions made in the linkage process should be carefully considered and documented, in particular the choice of the 'spine dataset', the definition of eligibility criteria and the point at which eligibility criteria are applied.
- Efficiency of spine linkage depends on high case ascertainment and data quality of linkage variables in the spine dataset. These aspects need to be carefully evaluated before the spine approach is used to create the analysis cohort.

Introduction

Using data linkage to combine information from records in separate data sources can provide a more detailed picture of characteristics of patients, their disease, the care they receive and their outcomes. For example, for patients undergoing emergency surgery for bowel cancer, information on patient, tumour and treatment characteristics can come from a clinical disease-specific dataset, information on emergency surgery from a clinical treatment-specific dataset and information on admissions and outcomes from a routinely collected administrative hospital dataset.

Methods for linking two datasets are well established.[1–3] However, when linking more than two datasets, many decisions need to be made (Table 1), including which datasets to link together.[4] 'Pairwise linkages' (i.e. linking each dataset to every other dataset) offer the most inclusive approach.[5] However, the number of linkages quickly escalates with the number of datasets that need to be linked (Supplementary Table A1, available as Supplementary data at *IJE* online), which can add delays, increase costs and require transfer of personal information between multiple organizations. An alternative approach is to treat one dataset as the 'spine dataset' and link each of the other datasets to this spine. For example, four datasets can be combined using three linkages in the spine approach (Supplementary Figure A1, available as Supplementary data at *IJE* online), whereas the pairwise approach would use six linkages (Supplementary Figure A2, available as Supplementary data at *IJE* online).

If the spine dataset captures 100% of eligible patients and there is perfect linkage between all datasets, then the spine and pairwise approaches will be equivalent. In practice, few datasets have complete case ascertainment and missing or incorrect patient identifiers can lead to incomplete linkage.[6,7]

The spine approach has a number of potential limitations. First, patients who are missing from the spine dataset, or not linked to the spine dataset, cannot be included in the analysis. In addition, records in non-spine datasets can only be identified as belonging to the same patient if records link indirectly via the spine dataset. Consequently, the spine approach will in general lead to a smaller analysis cohort, which may affect how well the analysis dataset represents the full population. That is, if some patient groups are less likely to be recorded in some datasets, spine linkage will suffer from selection bias. Conversely, although pairwise linkage is more inclusive, individual data items may have more missing values due to the inclusion of more patients who do not appear in all datasets. Thus spine linkage may seem to have more complete data than pairwise linkage.

Our aim was to compare spine and pairwise linkage using a real-world example of patients undergoing emergency bowel cancer surgery, with data from an administrative hospital dataset and two clinical datasets. We compared approaches by considering the number of eligible patients linked by each approach, characteristics of these patients, levels of missing data and whether analysis results were sensitive to the approach used.

Methods

Spine approach vs pairwise approach to linkage

We compared the spine and pairwise approaches, illustrated using three datasets: A, B and C, where A represents the spine dataset (Figure 1). In the spine approach, A is linked to B and to C (the non-spine datasets) separately, with no direct link between datasets B and C. Records can then be classified into six subgroups represented as rows of blocks in Figure 1, defined by whether there was linkage between datasets A and B, datasets A and C or both. For example, Row 1 represents records in A that did not link to B or C, whereas Row 4 represents records that linked between A and B and between A and C. The pairwise approach uses all three pairwise linkages (A to B, A to C and

Table 1 Decisions to be made when linking multiple datasets

Decisions to be made:	Available options include:	In this example, we used:
Choice of linkage methods between pairs of datasets	Deterministic linkage Probabilistic linkage Combination of both	Deterministic linkage
Strategy for which datasets to link together	Spine approach Pairwise approach (Other approaches)	Comparison of spine and pairwise approaches
Selection of linkage variables	Desired characteristics: objective (e.g. administrative rather than clinical), good completeness, available in at least two datasets Contribution to probabilistic linkage can be quantified with respect to data quality and chance agreement[2]	NHS number, sex, date of birth, residential postcode (used in deterministic linkage carried out by trusted third party)[1,22]
Selection of analysis cohort	Depends on the research question, linkage strategy used and the data source that includes outcomes	For spine approach: maximum analysis cohort is patients in spine dataset For pairwise approach: maximum analysis cohort is patients in any dataset
Reconciling information when available from more than one source	Context-dependent Use expert knowledge to guide rules for reconciling information	In general, clinical datasets take precedence over administrative (spine) dataset Details in Supplementary material, Section C, available as Supplementary data at *IJE* online
Dealing with incomplete data within a data source	Complete case analysis 'Ad hoc' missing data methods (e.g. missing indicator) Multiple imputation Use clinical knowledge to understand why data are missing	Complete case analysis

NHS, National Health Service.

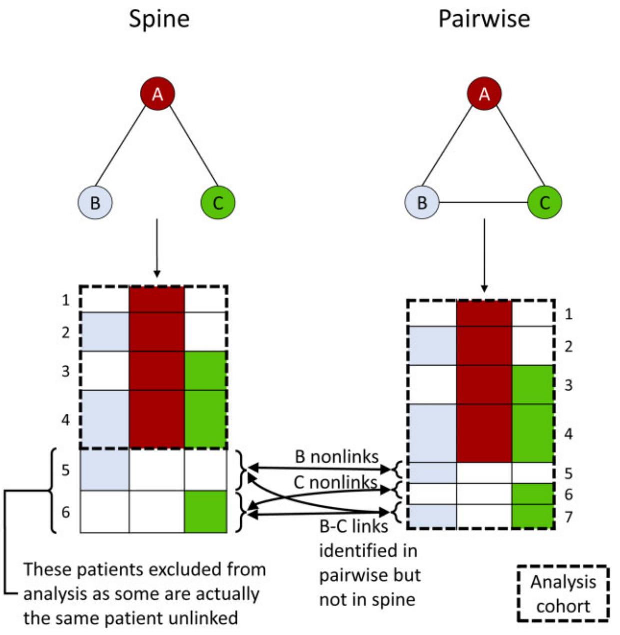
Spine
Pairwise
A
B
C
A
B
C
1
2
3
4
5
6
1
2
3
4
5
6
7
B nonlinks
C nonlinks
B-C links identified in pairwise but not in spine
These patients excluded from analysis as some are actually the same patient unlinked
Analysis cohort

Figure 1 Illustration of spine linkage vs pairwise linkage. Classification of subgroups: Both: 1—unlinked A records; 2—records linked between A and B; 3—records linked between A and C; 5—unlinked B records; 6—unlinked C records. Spine (left): 4—records linked between A and B, and between A and C. Pairwise (right): 4—records linked between A, B and C; 7—records linked between B and C.

B to C) leading to seven subgroups, the additional subgroup (Row 7) being those that linked between B and C but not to A.

Figure 1 illustrates that with spine linkage, the same individual may appear in the unlinked part of dataset B as well as in the unlinked part of dataset C (Subgroups 5 and 6 in Figure 1) because there has been no attempt to directly link dataset B to dataset C. This means that we may double count these individuals, who appear to be two distinct people rather than two records belonging to the same person. As a result of this duplication of records, the total size of the six subgroups in the spine approach (Figure 1, left panel) may appear to be greater than the total size of the seven subgroups in the pairwise approach (Figure 1, right panel). A solution is to exclude these unlinked subgroups from the spine linkage analysis cohort (bold dashed box) to avoid including the same individual twice. In contrast, the pairwise approach allows use of the direct linkage between B and C to identify which records in datasets B and C belong to the same individual, provided case ascertainment and linkage quality are high. If so, this reduces the risk of including the same individual twice and therefore the analysis cohort created by pairwise linkage can reasonably include all seven subgroups.

Data sources for patients undergoing emergency bowel cancer surgery

As a real-world example, we used three national datasets including patients who had emergency bowel cancer surgery in the English National Health Service (NHS). Clinical information on patients diagnosed with bowel cancer is contained in the disease-specific dataset collected by the National Bowel Cancer Audit (NBOCA), including information on patient and tumour characteristics, processes of care and health outcomes.[8] Clinical information about patients undergoing emergency bowel surgery is available from the procedure-specific National Emergency Laparotomy Audit (NELA), including information on physiological characteristics of patients, surgery and health outcomes.[9] Administrative information on all hospital episodes in the English NHS can be obtained from Hospital Episode Statistics (HES), collected for reimbursement purposes.[10,11] Each dataset contained information on mortality, provided by the Office of National Statistics.[12]

Linkage was carried out for NBOCA records in which the date of surgery was between 31 October 2013 and 30 April 2018, NELA records in which the admission date was between 1 December 2013 and 30 November 2019, and HES records for patients with a bowel cancer diagnosis or a bowel surgery procedure in any hospital episode between 31 October 2013 and 30 April 2018 (Supplementary Table B1, available as Supplementary data at *IJE* online). We used the maximum date range possible for each dataset during linkage in order to prevent missed links that could arise from applying restrictions prior to linkage.

Sources of each data item are given in Tables 2–4. The Index of Multiple Deprivation (IMD) is an area-based measure of socio-economic deprivation across seven domains, based on an area of residence typically including ∼1500 people and 650 households.[13] Patients were grouped into five categories based on quintiles of the national ranking of the IMD, where 1 represents the most deprived quintile and 5 represents the least deprived quintile. The American Society of Anesthesiologists (ASA) grade categorizes a patient's physical status from 1 (healthy) to 5 (moribund).[14] The performance status categorizes functional ability from 0 (normal activity) to 4 (no self-care).[15] Surgical urgency was defined according to the National Confidential Enquiry into Patient Outcome and Death Classification of Intervention 2014.[16,17] Diagnostic information used the International Statistical Classification of Diseases and Health Related Problems tenth revision (ICD-10) codes,[18] which were categorized by cancer site, and surgical procedure used the Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4 (OPCS-4) codes.[19] Cancer stage in four categories was derived from the final pathology Tumour, Node, Metastasis (TNM) staging in NBOCA[20] and from the level of malignancy based on surgical findings in NELA.[16] The number of co-morbidities was defined using ICD-10 codes in HES according to the Royal College of Surgeons of England Charlson Score.[21] Thirty-day unplanned readmission was defined as an emergency admission to any hospital for any cause within 30 days of surgery, according to HES.

To reconcile conflicting information for the same patient from different datasets, our guiding principles were to use the treatment-specific dataset as the preferred source of data about patients and their surgery, the disease-specific dataset as the preferred source of data about their bowel cancer and the administrative hospital dataset as the preferred source of administrative items, including mortality (Supplementary material, Section C, available as Supplementary data at *IJE* online).

Data linkage and analysis

For the spine approach, we used the administrative dataset HES as the spine dataset because it is expected to have good case ascertainment and data completeness.[10] For both approaches, linkage was undertaken using deterministic (i.e. rule-based) methods. For linkages with the spine dataset, pairs of records were considered linked if there was exact agreement on direct patient identifiers (the patients' unique NHS number, sex, date of birth and residential postcode).[1,22] For linkage between the non-spine datasets, pairs of records were considered linked if they matched on NHS number.

For both approaches, linkage was carried out on all available data. Thereafter, patients were retained for analysis if they underwent emergency surgery for bowel cancer in at least one dataset according to eligibility criteria (Supplementary Table B2, available as Supplementary data at *IJE* online). Since eligibility criteria were applied after linkage, we did not expect all patients to link across all three datasets, e.g. not all patients undergoing emergency surgery are patients with bowel cancer and not all bowel cancer patients undergo emergency surgery.

Comparing the spine and pairwise approaches

First, we compared patient numbers in the analysis cohorts created by spine and pairwise linkage. Second, we described characteristics of eligible patients captured by (i) spine approach, (ii) pairwise approach and (iii) pairwise approach but not spine approach. Proportions of patients with missing data were reported separately to patients with information not available due to incomplete linkage [i.e. not linked to the dataset(s) containing the relevant information]. Third, we compared unadjusted regression estimates of patient and tumour characteristics with mortality (logistic regression for 90-day, Cox regression for 2-year), complications (logistic regression) and length of stay (linear regression) according to the linkage approach. Each analysis included only patients with complete data on the outcome and covariates of interest.

In both linkage approaches, a decision must be made regarding when to apply eligibility criteria. In the main analysis, we undertook linkage on the full data available and then applied eligibility criteria. To reflect situations in which analysts request an extract of a dataset according to specified eligibility criteria, we conducted a sensitivity analysis in which broad eligibility criteria were applied before linkage and further eligibility criteria were applied after linkage (Supplementary Table B3, available as Supplementary data at *IJE* online).

Results

Numbers of patients in the analysis cohorts created by spine vs pairwise linkage

Spine linkage created an analysis cohort of 15 826 patients compared with 16 100 when pairwise linkage was used (Figure 2). Just over half of patients included in either linkage approach (8526/15 826 patients with spine and 8628/16 100 patients with pairwise) linked across all three datasets. For both linkage approaches, most patients (>95%) in the analysis cohort linked between at least two datasets. The spine analysis cohort was a subset of the pairwise cohort. The total numbers of eligible patients linked to the spine dataset (i.e. captured inside the HES circle of the Venn diagrams) differs between approaches because for some patients the additional linkage between the two non-spine datasets creates indirect links between the spine dataset and the non-spine datasets. See Supplementary Figure D1 (available as Supplementary data at *IJE* online) for further explanation.

Characteristics of the analysis cohorts created by spine vs pairwise linkage

Characteristics of patients included in the spine and pairwise analysis cohorts were almost identical (Tables 2–4) because the sizes of the cohorts were so similar. Proportions of missing data were also very similar. Note that Tables 2–4 are split into sections defined by how many datasets contribute to each variable. For example, the variable age in Table 2 comes from HES, with missing values imputed based on entries in the other two datasets, according to a pre-defined rule (see Supplementary material, Section C, available as Supplementary data at *IJE* online for details).

Characteristics of patients linked by pairwise linkage but not spine linkage

Of 274 additional patients captured in the pairwise approach (Figure 2), approximately two-thirds were only in the treatment-specific dataset (NELA) and one-third were only in the disease-specific dataset (NBOCA). Overall, the additional patients were more likely to have ASA Grade 3, rectal cancer and cancer stage 1–2 compared with the remaining patients in the pairwise analysis cohort, but other patient characteristics and processes of care were similar (Supplementary Table E1, available as Supplementary data at *IJE* online). Proportions of missing/unavailable data in performance status, cancer site and deprivation were markedly higher in the additional patients (71%, 65% and 65%, respectively) compared with the whole pairwise analysis cohort (29%, 4% and 2%). Mortality was lower in the additional patients, but they had more missing outcome data (32% vs 1%).

Comparison of unadjusted regression results for spine vs pairwise linkage

With such similar numbers in the two approaches, associations between patient and tumour characteristics and outcomes were not sensitive to the linkage approach (Figure 3, further detail in Supplementary Tables F1–F5, available as Supplementary data at *IJE* online). For these complete case analyses, each unadjusted regression analysis used data from >93% of the full analysis cohort for all patient and tumour characteristics and outcomes, except for unplanned return to theatre, which was complete for 63% of patients in both linkage approaches.

Sensitivity analysis results

A sensitivity analysis applying broad eligibility criteria before linkage (Supplementary Table B3, available as Supplementary data at *IJE* online) resulted in 14 509 patients in the spine approach cohort compared with 16 116 in the pairwise approach (Supplementary Figure G1, available as Supplementary data at *IJE* online); 1607 patients linked via pairwise but not spine linkage. The characteristics of patients in the spine and pairwise analysis cohorts were almost identical although mortality was slightly lower in the spine cohort, e.g. 2-year mortality: 15.2% in spine cohort vs 17.1% in pairwise

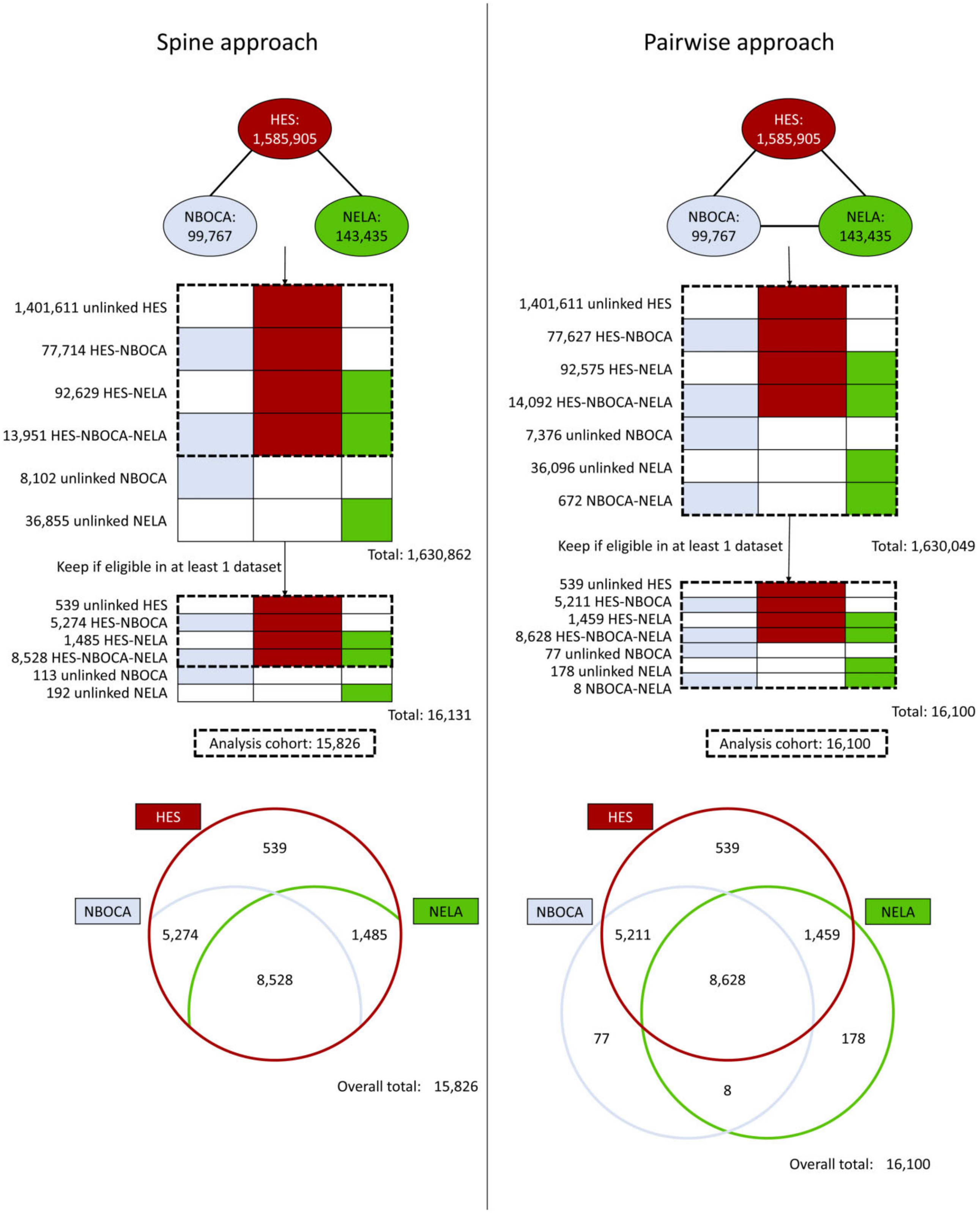

Figure 2 Linkage process and resulting Venn diagrams for spine linkage vs pairwise linkage. HES, Hospital Episode Statistics; NBOCA, National Bowel Cancer Audit; NELA, National Emergency Laparotomy Audit.

Table 2 Number of cases (percentage of those with complete data) for patient and tumour characteristics, processes of care and patient outcomes available in all three datasets, comparing analysis cohorts after spine linkage and pairwise linkage

		Spine approach		Pairwise approach	
		n	%	*n*	%
		(Total = 15 826)		(Total = 16 100)	
Available in all three datasets					
Age (years)	<50	1404	8.9	1445	9.0
	50–59	2130	13.5	2167	13.5
	60–74	5788	36.6	5893	36.7
	75–84	4676	29.6	4738	29.5
	≥85	1804	11.4	1833	11.4
	Missing (% of total)	24 (0.2)		24 (0.1)	
Sex	Female	7656	48.4	7793	48.4
	Male	8170	51.6	8306	51.6
	Missing (% of total)	0 (0.0)		1 (0.0)	
Surgical procedure	Colectomy: left/sigmoid/anterior resection	2349	14.8	2381	14.8
	Colectomy: right/ileocaecal	7853	49.6	7992	49.6
	Colectomy: subtotal/panprocto	1282	8.1	1306	8.1
	Hartmann	3209	20.3	3273	20.3
	Other resection: transverse/abdominoperineal resection of rectum/pelvic exenteration	465	2.9	473	2.9
	Stoma or other surgery	668	4.2	675	4.2
	Missing (% of total)	0 (0.0)		0 (0.0)	
Calendar year of surgical procedure	2013/2014	3772	23.8	3799	23.6
	2015	3466	21.9	3487	21.7
	2016	3767	23.8	3862	24.0
	2017/2018	4818	30.4	4938	30.7
	Missing (% of total)	3 (0.0)		14 (0.1)	
90-day mortality	Alive	14 335	90.6	14 509	90.6
	Dead	1487	9.4	1499	9.4
	Missing (% of total)	4 (0.0)		92 (0.6)	
2-year mortality	Alive	13 181	83.3	13 349	83.4
	Dead	2641	16.7	2659	16.6
	Missing (% of total)	4 (0.0)		92 (0.6)	

The number of records with missing data is given after each covariate has been summarized.

(Supplementary Table G1, available as Supplementary data at *IJE* online). The additional patients were more likely to have less advanced cancer, longer hospital stays and much higher mortality (Supplementary Table G1, available as Supplementary data at *IJE* online). Despite the differences in mortality, this had no impact on associations between baseline characteristics and outcomes statistics (Supplementary Figure G2, available as Supplementary data at *IJE* online).

Discussion

Summary

We considered differences between spine and pairwise linkage of three datasets, demonstrating how these approaches can be evaluated. In our example using real-world data, we found negligible differences in analysis cohorts created using spine or pairwise linkage. There were no systematic differences between patients linked using the two approaches, and associations between patient and tumour characteristics and outcomes were not sensitive to the linkage approach. Sensitivity analysis demonstrated the importance of applying eligibility criteria after spine linkage; if patients are identified as eligible in some datasets but not in others, applying strict eligibility criteria prior to linkage may result in missing links as well as different characteristics in unlinked patients, potentially leading to bias.

Strengths and limitations

Here, the analysis cohort created by spine linkage captured a very high proportion of patients included with the

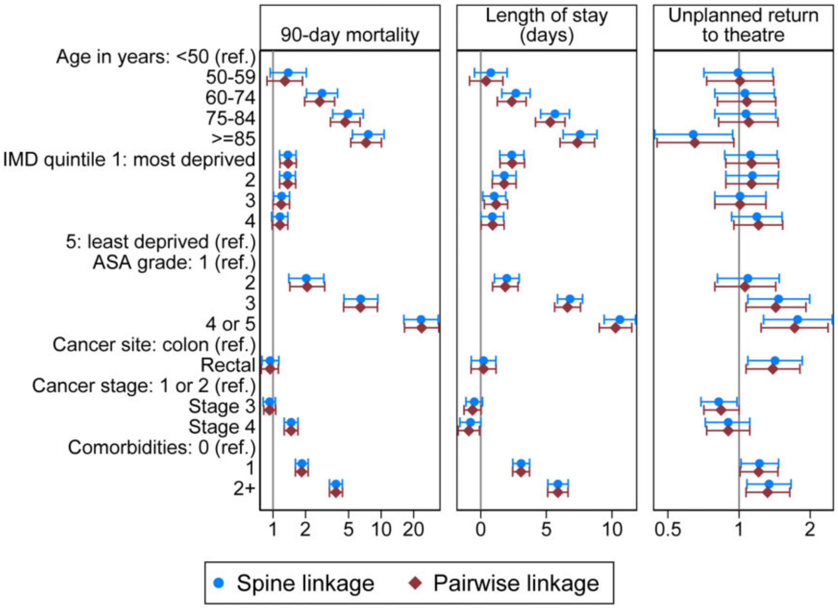

Figure 3 Unadjusted regression estimates and 95% confidence intervals for 90-day mortality (odds ratios), length of stay (mean differences) and unplanned return to theatre (odds ratios), comparing patients linked via spine linkage vs pairwise linkage. Ref., reference category; IMD, Index of Multiple Deprivation; ASA, American Society of Anesthesiologists.

pairwise approach. However, this should not be assumed to be the case in general. Performance of spine linkage was excellent here because the chosen spine dataset (HES) captured nearly all surgical patients treated in the English NHS, resulting in very high case ascertainment. Also, linkage error was low because of the availability of a common set of patient identifiers throughout the care pathway that are largely complete in all datasets.[2]

Where datasets arise from different systems, the choice of the spine dataset may not be obvious and linkage errors may be more common. For example, in a study linking paediatric critical care data to laboratory surveillance data, linkage errors were relatively common due to poor recording of identifiers.[23] Another study, which explored premature mortality in people with serious mental illnesses, recommended using both hospital care data and primary care data for case ascertainment after finding ascertainment bias in previous studies that used a single data source.[24]

The additional patients in the pairwise cohort who were not linked to the spine dataset typically had higher proportions of missing or unavailable data. Since this was a relatively small group, not including these patients had a negligible impact on observed associations of patient and tumour characteristics with outcomes in the spine approach.

A limitation of the spine approach is that in a study in which the outcome is defined by linkage, even a small proportion of missed links could lead to ascertainment bias.[25,26] Missed links can lead to underestimation of outcomes captured in the linked data, which is problematic when this occurs differentially according to variables of interest. For example, a Canadian study linking administrative datasets to immigration and mortality data found lower linkage rates for people born in East Asia and for some causes of death.[27]

The spine approach does not allow identification of missed links between non-spine datasets (NBOCA and NELA in our example). However, it should be noted that even if direct linkage between non-spine datasets was available, as in the pairwise approach, missed links could still occur as no linkage process is perfect.[25]

This study used deterministic methods to link datasets. Probabilistic linkage methods could have been used to reduce linkage error.[2,28] However, given the negligible difference between the spine and pairwise cohorts here, it is unlikely that probabilistic linkage would have an impact on findings. Furthermore, if probabilistic linkage were to be incorporated into a pairwise approach, adding this further complexity to an already computationally intensive process may negate any gains.[29]

When information was missing from one dataset but available in one (or more) of the other datasets, linkage allowed us to reduce the amount of missing data by ‘recovering’ this information from one of the other datasets. Consequently, there was very low missing data in the analysis cohorts and complete case analysis could be

Table 3 Number of cases (percentage of those with complete data) for patient and tumour characteristics, processes of care and patient outcomes available in two datasets only, comparing analysis cohorts after spine linkage and pairwise linkage

		Spine approach		Pairwise approach	
		n (Total = 15 826)	%	*n* (Total = 16 100)	%
Available in two datasets only					
IMD quintile (HES, NBOCA)	1: most deprived	2719	17.3	2746	17.4
	2	2972	19.0	2992	19.0
	3	3215	20.5	3232	20.5
	4	3398	21.7	3422	21.7
	5: least deprived	3373	21.5	3381	21.4
	Missing (% of total)	149 (0.9)		149 (0.9)	
	Unavailable (% of total)	0 (0.0)		178 (1.1)	
ASA grade (NBOCA, NELA)	1	1670	11.3	1694	11.3
	2	6244	42.4	6359	42.4
	3	5242	35.6	5343	35.6
	4 or 5	1581	10.7	1616	10.8
	Missing (% of total)	550 (3.5)		549 (3.4)	
	Unavailable (% of total)	539 (3.4)		539 (3.3)	
Cancer site (HES, NBOCA)	Colon	13 802	89.4	13 884	89.4
	Rectal	1632	10.6	1648	10.6
	Missing (% of total)	392 (2.5)		390 (2.4)	
	Unavailable (% of total)	0 (0.0)		178 (1.1)	
Cancer stage (NBOCA, NELA)	Stage 1 or 2	6271	42.3	6383	42.4
	Stage 3	5834	39.4	5909	39.2
	Stage 4	2717	18.3	2776	18.4
	Missing (% of total)	465 (2.9)		493 (3.1)	
	Unavailable (% of total)	539 (3.4)		539 (3.3)	
Length of stay (days) (HES, NELA)	0–7	4535	29.7	4569	29.5
	8–14	5525	36.2	5606	36.2
	15–21	2366	15.5	2415	15.6
	22–28	1060	6.9	1075	6.9
	>28	1791	11.7	1821	11.8
	Missing (% of total)	549 (3.5)		537 (3.3)	
	Unavailable (% of total)	0 (0.0)		77 (0.5)	

The number of 'missing' and 'unavailable' cases is given after each covariate has been summarized. 'Missing' refers to records in which there is linkage to the source(s) of the data item but the information is missing. 'Unavailable' refers to records in which there is no linkage to either source of the data item. IMD, Index of Multiple Deprivation; ASA, American Society of Anesthesiologists; HES, Hospital Episode Statistics; NBOCA, National Bowel Cancer Audit; NELA, National Emergency Laparotomy Audit.

used.[30] In general, careful consideration is needed to understand the reasons for missing data and why data items are not completed, including discussions with clinical colleagues and colleagues responsible for entering data. An alternative could have been to include all eligible patients and use missing data methods, such as multiple imputation.[30,31]

Implications

The key benefit of spine linkage compared with pairwise is that it is more time-efficient, resource-efficient and cost-efficient because fewer data linkages are required. Requiring fewer linkages also reduces risks of disclosure of sensitive information, thus enhancing data security. These benefits are likely to grow the more datasets there are to link together.

In order for the spine approach to be appropriate, the nominated spine dataset must have excellent case ascertainment.[32] Case ascertainment is usually high for datasets that capture major procedures or events and can be checked by considering proportions of eligible patients in each dataset who link to the spine dataset. Further work is needed to investigate the level of case ascertainment required in general. We also need low linkage error between pairs of datasets. This is likely to be true for datasets containing a unique patient identifier, such as the NHS number used in England.[6]

Table 4 Number of cases (percentage of those with complete data) for patient and tumour characteristics, processes of care and patient outcomes available in one dataset only, comparing analysis cohorts after spine linkage and pairwise linkage

		Spine approach		Pairwise approach	
		n (Total = 15 826)	%	*N* (Total = 16 100)	%
Available in one dataset only					
Co-morbidities (HES)	0	8020	53.0	8024	53.0
	1	4522	29.9	4526	29.9
	2+	2577	17.0	2578	17.0
	Missing (% of total)	707 (4.5)		709 (4.4)	
	Unavailable (% of total)	0 (0.0)		263 (1.6)	
Performance status (NBOCA)	Normal activity	4681	41.5	4710	41.3
	Walk and light work	3796	33.6	3843	33.7
	Walk and all self-care	1917	17.0	1938	17.0
	Limited or no self-care	897	7.9	903	7.9
	Missing (% of total)	2511 (15.9)		2530 (15.7)	
	Unavailable (% of total)	2024 (12.8)		2176 (13.5)	
Surgical urgency (NELA)	Expedited (>18 h)	2339	23.5	2394	23.4
	Urgent (6–18 h)	3940	39.5	4044	39.5
	Urgent (2–6 h)	2886	28.9	2967	29.0
	Immediate or emergency (<2 h, or resus of >2 h possible)	808	8.1	827	8.1
	Missing (% of total)	40 (0.3)		41 (0.3)	
	Unavailable (% of total)	5813 (36.7)		5827 (36.2)	
Emergency readmission within 30 days (HES)	No	13 613	90.0	13 621	90.0
	Yes	1506	10.0	1507	10.0
	Missing (% of total)	707 (4.5)		709 (4.4)	
	Unavailable (% of total)	0 (0.0)		263 (1.6)	
Unplanned return to theatre (NELA)	No	9237	93.4	9471	93.3
	Yes	658	6.6	679	6.7
	Missing (% of total)	118 (0.7)		123 (0.8)	
	Unavailable (% of total)	5813 (36.7)		5827 (36.2)	

The number of 'missing' and 'unavailable' cases is given after each covariate has been summarized. 'Missing' refers to records in which there is linkage to the source(s) of the data item but the information is missing. 'Unavailable' refers to records in which there is no linkage to either source of the data item. HES, Hospital Episode Statistics; NBOCA, National Bowel Cancer Audit; NELA, National Emergency Laparotomy Audit.

Suitability of the spine approach also depends on the research question. For example, in effectiveness research, we analyse linked records to produce unbiased estimates of exposure–outcome relationships. However, if we were estimating absolute levels of an outcome, we would need these estimates to be unbiased. For example, in healthcare performance assessment, between-hospital variation in linkage rates to the spine dataset could affect comparisons of performance indicators among hospitals.[33]

Sensitivity analysis demonstrated that the analysis cohort created using spine linkage depended on when eligibility criteria were applied. If eligibility criteria are applied when defining the datasets to be linked (e.g. when requesting data extracts to be linked), the spine approach may not be appropriate: there may be missing links if patients are identified as eligible in some of the datasets but not in others, resulting in the spine approach capturing fewer eligible patients. Also, there may be substantial differences in characteristics of those not linked via the spine approach, potentially leading to bias, particularly in settings with low case ascertainment or a higher rate of linkage errors (e.g. missing data on personal identifiers).

In general, when choosing the spine dataset, factors to consider include ascertainment of the population of interest, and availability and completeness of linkage variables. We chose an administrative dataset that is used for reimbursement purposes[10] and thus case ascertainment and data completeness were high. However, in different settings, administrative datasets may not be the optimal choice of spine dataset. For example, in a study considering the linkage of routine birth records, the administrative hospital admissions dataset had poor case ascertainment compared with national birth registration records.[34] In some cases, the most useful spine option might be an 'independent' population spine, i.e. a dataset of identifiers capturing the entirety of the relevant population but not containing any variables required for the analysis. For example, the Personal Demographic Service (a database of identifiers for all individuals with an NHS number held by NHS Digital) has been used to in England to facilitate linkage between non-health datasets (specifically, the National Pupil Database) and HES.[35] A similar approach is taken to linking multi-agency data in Australia.[36]

In practice, the pairwise approach may not always be feasible. In that case, the spine approach can only be validated using generic methods for assessing linkage quality: comparing patient characteristics, care processes and patient outcomes between patients linked and not linked to the spine dataset; and investigating unlikely or implausible links and unlinked records that were expected to link.[7,37,38]

Conclusion

We demonstrate that spine linkage can be used as an efficient alternative to pairwise linkage. The spine approach requires fewer linkages between pairs of datasets, thus reducing delays, costs and resources needed and increasing data security. However, researchers should systematically evaluate case ascertainment and potential for linkage error in the nominated spine dataset before spine linkage is used to create the analysis cohort.

Ethics approval

As the National Bowel Cancer Audit involves analysis of data for service evaluation, it is exempt from UK National Research Ethics Committee approval. Section 251 approval was obtained from the Ethics and Confidentiality Committee for the collection of personal health data without the consent of patients. The study was performed in accordance with the Declaration of Helsinki.

Data availability

The data used in this study are available from NHS Digital and Public Health England's Office for Data Release but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. We do not have permission to share the patient-level records used in our analysis.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

H.B.: data curation, formal analysis, methodology, writing of original draft, review and editing. L.S.: funding acquisition, methodology, writing of original draft, review and editing. K.H.: funding acquisition, methodology, writing of original draft, review and editing. J.v.d.M.: conceptualization, methodology, funding acquisition, writing of original draft, review and editing. K.W.: conceptualization, methodology, funding acquisition, writing of original draft, review and editing.

Funding

This study/project is funded by the National Institute for Health Research (NIHR) Health Service and Delivery Research Programme (Grant 17/05/45). This report is independent research supported by the NIHR ARC North Thames. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. K.H. is funded by the Wellcome Trust (Grant 212953/Z/18/Z). This research was supported in part by the NIHR Great Ormond Street Hospital Biomedical Research Centre and the Health Data Research UK (grant no. LOND1), which is funded by the UK Medical Research Council and eight other funders.

Acknowledgements

The National Bowel Cancer Audit is commissioned by the Healthcare Quality Improvement Partnership (HQIP) as part of the National Clinical Audit and Patient Outcomes Programme, and funded by NHS England and the Welsh Government (www.hqip.org.uk/national-programmes). Neither HQIP nor the funders had any involvement in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The National Emergency Laparotomy Audit is commissioned by the HQIP as part of the National Clinical Audit and Patient Outcomes Programme (NCAPOP). The programme is funded by NHS England, the Welsh Government and, with some individual projects, other devolved administrations and crown dependencies: hqip.org.uk/national-programmes. This work uses data provided by patients and collected by the NHS as part of their care and support.

Conflict of interest

None declared.

References

1. Harron K, Mackay E, Elliot M. An introduction to data linkage: Administrative Data Research Network. 2016. http://eprints.ncrm.ac.uk/4282/ (2 November 2020, date last accessed).
2. Blake HA, Sharples LD, Harron K, van der Meulen JH, Walker K. Probabilistic linkage without personal information successfully linked national clinical datasets. *J Clin Epidemiol* 2021;**136**:136–45.
3. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *J Biomed Inform* 2015;**56**:80–86.
4. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Ann Hum Biol* 2020;**47**:218–26.

5. Sadinle M, Fienberg SE. A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. *J Am Stat Assoc* 2013;**108**:385–97.
6. Harron K, Dibben C, Boyd J *et al.* Challenges in administrative data linkage for research. *Big Data Soc* 2017;**4**: 2053951717745678.
7. Gilbert R, Lafferty R, Hagger-Johnson G *et al.* GUILD: GUidance for Information about Linking Data sets. *J Public Health (Oxf)* 2018;**40**:191–98.
8. National Bowel Cancer Audit. Annual Report 2019. www.nboca.org.uk/reports/annual-report-2019/ (31 March 2020, date last accessed).
9. National Emergency Laparotomy Audit. The Sixth Patient Report of the NELA. 2020. https://www.nela.org.uk/Sixth-Patient-Report (9 November 2021, date last accessed).
10. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017;**46**:1093.i.
11. NHS Digital. Hospital Episode Statistics (HES). 2019. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (25 May 2020, date last accessed).
12. Office for National Statistics. Deaths registered in England and Wales. 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsregisteredinenglandandwalesseriesdrreferencetables (1 December 2020, date last accessed).
13. Ministry of Housing, Communities & Local Government. English indices of deprivation. 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (14 September 2020, date last accessed).
14. Daabiss M. American Society of Anaesthesiologists physical status classification. *Indian J Anaesth* 2011;**55**:111–15.
15. Oken MM, Creech RH, Tormey DC *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* 1982;**5**:649–56.
16. National Emergency Laparotomy Audit. Participant Manual. 2015. https://www.nela.org.uk/downloads/National Emergency Laparotomy Audit—Participant Manual—version 1.6.pdf (21 December 2021, date last accessed).
17. National Confidential Enquiry into Patient Outcome and Death. The NCEPOD Classification of Intervention. 2004. https://www.ncepod.org.uk/classification.html (1 December 2020, date last accessed).
18. NHS Digital. International Statistical Classification of Diseases and Health Related Problems (ICD-10) 5th Edition. 2018. https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/scci0021-international-statistical-classification-of-diseases-and-health-related-problems-icd-10-5th-edition (24 September 2019, date last accessed).
19. NHS Digital. OPCS Classification of Interventions and Procedures. 2020. https://datadictionary.nhs.uk/supporting_information/opcs_classification_of_interventions_and_procedures.html (2 November 2020, date last accessed).
20. Colorectal cancer staging. *CA Cancer J Clin* 2004;**54**:362–65.
21. Armitage JN, van der Meulen JH; Royal College of Surgeons Co-morbidity Consensus Group. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg* 2010;**97**:772–81.
22. Paixão ES, Harron K, Andrade K *et al.* Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in Brazil. *BMC Med Inform Decis Mak* 2017;**17**:108.
23. Harron K, Goldstein H, Wade A, Muller-Pebody B, Parslow R, Gilbert R. Linkage, evaluation and analysis of national electronic healthcare data: application to providing enhanced bloodstream infection surveillance in paediatric intensive care. *PLoS One* 2013;**8**:e85278.
24. John A, McGregor J, Jones I *et al.* Premature mortality among people with severe mental illness: new evidence from linked primary care data. *Schizophr Res* 2018;**199**:154–62.
25. Bohensky MA, Jolley D, Sundararajan V *et al.* Data Linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010;**10**:346.
26. Hagger-Johnson G, Harron K, Fleming T *et al.* Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open* 2015;**5**:e008118.
27. Chiu M, Lebenbaum M, Lam K *et al.* Describing the linkages of the immigration, refugees and citizenship Canada permanent resident data and vital statistics death registry to Ontario's administrative health database. *BMC Med Inform Decis Mak* 2016;**16**: 1–11.
28. Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linking to enhance deterministic algorithms and reduce linkage errors in hospital administrative data. *BMJ Health Care Inform* 2017;**24**:234–46.
29. Doidge JC, Harron K. Demystifying probabilistic linkage: common myths and misconceptions. *Int J Popul Data Sci* 2018;**3**: 410.
30. Lee KJ, Tilling KM, Cornish RP *et al.*; STRATOS initiative. Framework for the treatment and reporting of missing data in observational studies: the treatment and reporting of missing data in observational studies framework. *J Clin Epidemiol* 2021; **134**:79–88.
31. Sterne JAC, White IR, Carlin JB *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;**338**:b2393.
32. Black A. The IDI prototype spine's creation and coverage. Statistics New Zealand Working Paper No. 16–03. 2016. http://archive.stats.govt.nz/methods/research-papers/working-papers-original/idi-prototype-spine.aspx (24 November 2020, date last accessed).
33. Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. Utilising identifier error variation in linkage of large administrative data sources. *BMC Med Res Methodol* 2017;**17**:23–29.
34. Murray J, Saxena S, Modi N *et al.*; Medicines for Neonates Investigator Group. Quality of routine hospital birth records and the feasibility of their use for creating birth cohorts. *J Public Health (Oxf)* 2013;**35**:298–307.
35. Libuy N, Harron K, Gilbert R, Caulton R, Cameron E, Blackburn R. Linking education and hospital data in England: linkage process and quality. *Int J Popul Data Sci* 2021;**6**:1671.

36. Frazer B. Person spine linkage methodology and maintenance. *Int J Popul Data Sci* 2020;**5**:1566.
37. Doidge J, Christen P, Harron K, Quality assessment in data linkage. Office for National Statistics and Government Analysis Function. 2020. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage (28 August 2020, date last accessed).
38. Harron KL, Doidge JC, Knight HE *et al.* A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol* 2017;**46**:1699–710.